Prediction of Protein-Protein Interaction Strength Using Domain Features with Supervised Regression

Kamada, Mayumi; Sakuma, Yusuke; Hayashida, Morihiro; Akutsu, Tatsuya

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 240673, 7 pages
http://dx.doi.org/10.1155/2014/240673

Hindawi

Research Article

Prediction of Protein-Protein Interaction Strength Using Domain Features with Supervised Regression

Mayumi Kamada,[1] Yusuke Sakuma,[2] Morihiro Hayashida,[3] and Tatsuya Akutsu[3]

[1] *Department of Biosciences and Informatics, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama 223-8522, Japan*
[2] *Japan Ichiba Section Development Unit, Rakuten Inc., 4-12-3 Higashi-shinagawa, Shinagawa-ku, Tokyo 140-0002, Japan*
[3] *Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan*

Correspondence should be addressed to Morihiro Hayashida; morihiro@kuicr.kyoto-u.ac.jp

Received 3 April 2014; Accepted 30 May 2014; Published 24 June 2014

Academic Editor: Loris Nanni

Proteins in living organisms express various important functions by interacting with other proteins and molecules. Therefore, many efforts have been made to investigate and predict protein-protein interactions (PPIs). Analysis of strengths of PPIs is also important because such strengths are involved in functionality of proteins. In this paper, we propose several feature space mappings from protein pairs using protein domain information to predict strengths of PPIs. Moreover, we perform computational experiments employing two machine learning methods, support vector regression (SVR) and relevance vector machine (RVM), for dataset obtained from biological experiments. The prediction results showed that both SVR and RVM with our proposed features outperformed the best existing method.

1. Introduction

In cellular systems, proteins perform their functions by interacting with other proteins and molecules, and protein-protein interactions (PPIs) play various important roles. Therefore, revealing PPIs is a key to understanding biological systems, and many investigations and analyses have been done. In addition, a variety of computational methods to predict and analyze PPIs have been developed, for example, methods for predicting PPI pairs using only sequences information [1–5], for predicting amino acid residues contributing to PPIs [6–8], and for assessing PPI reliability in PPI networks [9, 10]. As well as studies of PPIs, analyses of strengths of PPIs are important because such strengths are involved in functionality of proteins. In terms of transcription factor complexes, if a constituent protein has a weak binding affinity, target genes may not be transcribed depending on intracellular circumstance. For example, it is known that multi-subunit complex NuA3 in *Saccharomyces Cerevisiae* consists of five proteins, Sas3, Nto1, Yng1, Eaf6, and Taf30, acetylates lysine 14 of histone H3, and activates gene transcription. However, only Yng1 and Nto1 are found solely in the complex, and interaction strengths between each component protein are thought to be different and transient. Hence, Byrum et al. proposed a biological methodology for identifying stable and transient protein interactions recently [11].

Although many biological experiments have been conducted for investigating PPIs [12, 13], strengths of PPIs have not been always provided. Ito et al. conducted large-scale yeast two-hybrid experiments for whole yeast proteins. In their experiments, yeast two-hybrid experiments were conducted for each protein pair multiple times, the number of experiments that observe interactions, or the number of interaction sequence tags (ISTs), was counted. Consequently, they decided that protein pairs having three or more ISTs should interact and reported interacting protein pairs.

The ratio of the number of ISTs to the total number of experiments for a protein pair can be regarded as the interaction strength between their proteins. On the basis of this consideration, several prediction methods for strengths of PPIs have been developed. LPNM [14] is a linear programming-based method; ASNM [15] is a modified method from the association method [16] for predicting PPIs. Chen et al. proposed association probabilistic method

(APM) [17], which is the best existing method for predicting strengths of PPIs as far as we know.

These methods are based on a probabilistic model of PPIs and make use of protein domain information. Domains are known as structural and functional units in proteins and well-conserved regions in protein sequences. The information of domains is stored in several databases such as Pfam [18] and InterPro [19]. The same domain can be identified in several different proteins. In these prediction methods, interaction strengths between domains are estimated from known interaction strengths between proteins, and interaction strengths for target protein pairs are predicted from estimated strengths of domain-domain interactions (DDIs).

On the other hand, Xia et al. proposed a feature-based method using neural network with features based on constituent domains of proteins [20], and they compared their method with the association method and the expectation-maximization method [21]. For the feature-based prediction of PPI strengths, we also utilize domain information and propose several feature space mappings from protein pairs. We use supervised regression and perform threefold cross validation for dataset obtained from biological experiments. This paper augments the preliminary work presented in conference proceedings [22]. Specifically, major augmentations of this paper and differences from the preliminary conference version are summarized as follows.

(i) We employ two supervised regression methods: support vector regression (SVR) and relevance vector machine (RVM). Note that we used only SVR with the polynomial kernel in the preliminary version [22].

(ii) The Laplacian kernel is used as the kernel function for SVR and RVM, and kernel parameters are selected via fivefold cross validation.

(iii) We prepare the dataset from WI-PHI dataset [23] with high reliability.

The computational experiments showed that the average root mean square error (RMSE) by our proposed method was smaller than that by the best existing method, APM [17].

2. Materials and Methods

In this section, we briefly review a probabilistic model and related methods, and propose several feature space mappings using domain information.

2.1. Probabilistic Model of PPIs Based on DDIs. There are some computational prediction methods for PPI strengths, and they are based on the probabilistic model of PPIs proposed by Deng et al. [21]. This model utilizes DDIs and assumes that two proteins interact with each other if and only if at least one pair of the domains contained in the respective proteins interacts. Figure 1(a) illustrates an example of this interaction model. In this example, there are two proteins P_1 and P_2, which consist of domains D_1, D_2 and domains D_2, D_3, D_4, respectively. According to Deng's model, if P_1 and P_2 interact, at least one pair among (D_1, D_2), (D_1, D_3), (D_1, D_4), (D_2, D_2), (D_2, D_3), and (D_2, D_4) interacts. Conversely, if a pair, for instance, (D_2, D_4), interacts, P_1 and P_2 interact. From the assumption of this model, we can derive the following simple probability that two proteins P_i and P_j interact with each other:

$$\Pr\left(P_{ij} = 1\right) = 1 - \prod_{D_m \in P_i, D_n \in P_j} \left(1 - \Pr\left(D_{mn} = 1\right)\right), \quad (1)$$

where $P_{ij} = 1$ indicates the event that proteins P_i and P_j interact (otherwise, $P_{ij} = 0$), $D_{mn} = 1$ indicates the event that domains D_m and D_n interact (otherwise, $D_{mn} = 0$), and P_i and P_j also represent the sets of domains contained in P_i and P_j, respectively. Deng et al. applied the EM (expectation maximization) algorithm to the problem of maximizing log-likelihood functions, the estimated probabilities that two domains interact, $\Pr(D_{mn} = 1)$, and proposed a method for predicting PPIs using the estimated probabilities of DDIs [21]. Actually, they calculated $\Pr(P_{ij} = 1)$ using (1) and determined whether or not P_i and P_j interact by introducing a threshold θ; that is, P_i and P_j interact if $\Pr(P_{ij} = 1) \geq \theta$; otherwise, the proteins do not interact.

As Deng's method, typical PPIs prediction methods based on domains have the following two steps. First the interaction between domains contained in interacting proteins is inferred from existing protein interaction data. And then, an interaction between new protein pairs is predicted on the basis of the inferred domain interactions using a certain model. Figure 1(b) illustrates the flow of this type of PPIs prediction. Since interacting sites may not be always included in some known domain region, it can cause the decrease of prediction accuracy in this framework.

2.2. Association Method: Inferring DDI from PPI Data. As described previously, probability of PPIs could be predicted based on probabilities of DDIs. In this subsection, we will briefly review related methods to estimate a probability of interaction for domain pair.

2.3. Association Method. Let $\mathscr{P}$ be a set of protein pairs that have been observed to interact or not. The association method [16] gives the following simple score for two domains D_m and D_n using proteins that include the following domains:

$$\mathrm{ASSOC}\left(D_m, D_n\right) = \frac{\left|\left\{\left(P_i, P_j\right) \in \mathscr{P} \mid D_m \in P_i, D_n \in P_j, P_{ij} = 1\right\}\right|}{\left|\left\{\left(P_i, P_j\right) \in \mathscr{P} \mid D_m \in P_i, D_n \in P_j\right\}\right|}, \quad (2)$$

where $|S|$ indicates the number of elements contained in the set S. This score represents the ratio of the number of interacting protein pairs including D_m and D_n to the total number of protein pairs including D_m and D_n. Hence, it can be considered as the probability that D_m and D_n interact.

2.4. Association Method for Numerical Interaction Data (ASNM). Originally the association method has been designed for inferring binary protein interactions. To predict numerical interactions such as interaction strengths, Hayashida et al. proposed the association method for numerical interaction (ASNM) by the modification of the original

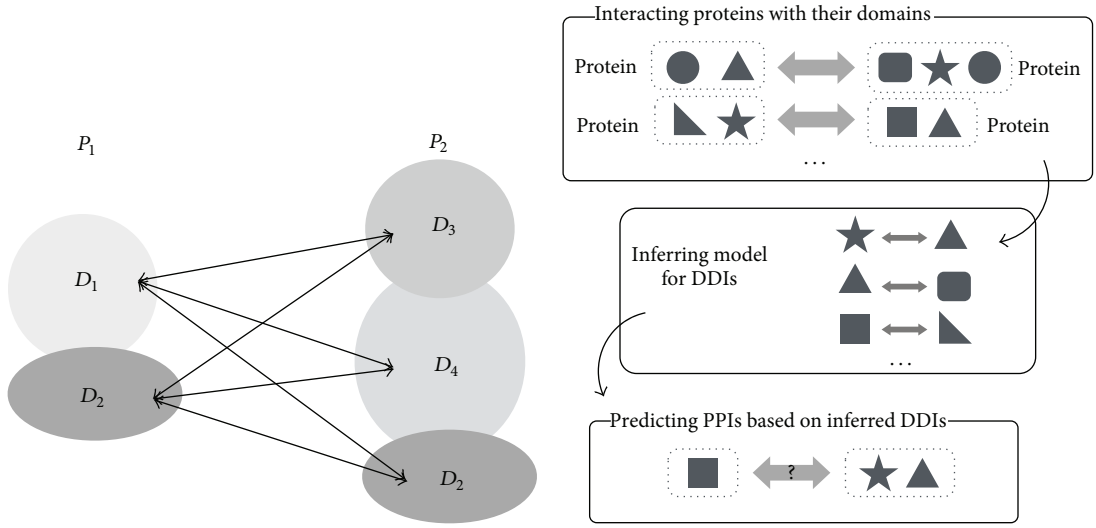

FIGURE 1: (a) Illustration of protein-protein interactions (PPIs) model based on domain-domain interactions (DDIs). (b) Schematic overview of PPIs prediction based on DDIs.

association method [15]. This method takes strengths of PPIs as input data. Let ρ_{ij} represent the interaction strength between P_i and P_j, and we suppose that ρ_{ij} is defined for all $(P_i, P_j) \in \mathcal{P}$. Then, the ASNM score for domains D_m and D_n is defined as the average strength over protein pairs including D_m and D_n by

$$\mathrm{ASNM}(D_m, D_n) = \frac{\sum_{\{(P_i,P_j)\in\mathcal{P} \mid D_m\in P_i, D_n\in P_j\}} \rho_{ij}}{\left|\left\{\left(P_i, P_j\right) \in \mathcal{P} \mid D_m \in P_i, D_n \in P_j\right\}\right|}. \tag{3}$$

If ρ_{ij} always takes only 0 or 1, $\mathrm{ASNM}(D_m, D_n)$ becomes $\mathrm{ASSOC}(D_m, D_n)$.

2.5. Association Probabilistic Method (APM). Although ASNM is a simple average of strengths of PPIs, Chen et al. proposed the association probabilistic method (APM) by replacing the strength with an improved strength [17]. It is based on the idea that the contribution of one domain pair to the strength of PPI should vary depending on the number of domain pairs included in a protein pair. They assumed that the interaction probability of each domain pair is equivalent in a protein pair, and transformed (1) as follows:

$$\Pr\left(D_{mn} = 1\right) = 1 - \left(1 - \Pr\left(P_{ij} = 1\right)\right)^{1/|P_i||P_j|}. \tag{4}$$

Thus, by substituting the numerator of ASNM, APM is defined by

$$\mathrm{APM}(D_m, D_n) = \frac{\sum_{\{(P_i,P_j)\in\mathcal{P} \mid D_m\in P_i, D_n\in P_j\}} \left(1 - \left(1 - \rho_{ij}\right)^{1/|P_i||P_j|}\right)}{\left|\left\{\left(P_i, P_j\right) \in \mathcal{P} \mid D_m \in P_i, D_n \in P_j\right\}\right|}. \tag{5}$$

They conducted some computational experiments, and reported that APM outperforms existing prediction methods such as ASNM and LPNM.

2.6. Proposed Feature Space Mappings from Protein Pairs. The association methods including ASNM and APM are based on the probabilistic model of PPIs defined by (1), and infer strengths of PPIs from estimated DDIs using given frequency of interactions or interaction strengths of protein pairs. On the other hand, we can also infer PPI strengths utilizing features obtained from given information such as sequence and structure of proteins with machine learning methods. Xia et al. proposed a method to infer strengths of PPIs using artificial neural network with features from constituent domains of proteins [20]. In this paper, for predicting strengths of PPIs, we propose several feature space mappings from protein pairs making use of domain information.

2.7. Feature Based on Number of Domains (DN). As described above, constituent domains information is useful for inferring PPIs and also can be used as a representation of each protein. Actually, Xia et al. represented each protein by binary numbers indicating whether a protein has a domain or not based on the information of constituent domains, and used them with the artificial neural network to predict PPI strengths [20]. Here, it can be considered that the probability that two proteins interact increases with a larger number of domains included in the proteins. Therefore, in this paper, we propose a feature space mapping based on the number of constituent domains (called DN) from two proteins. The feature vector of DN for two proteins P_i and P_j is defined by

$$\begin{aligned} f_{ij}^{(m)} &= M\left(D_m, P_i\right) \quad \text{for } D_m \in P_i, \\ f_{ij}^{(T+n)} &= M\left(D_n, P_j\right) \quad \text{for } D_n \in P_j, \\ f_{ij}^{(l)} &= 0 \quad \text{for } D_l \notin P_i \cup P_j, \end{aligned} \tag{6}$$

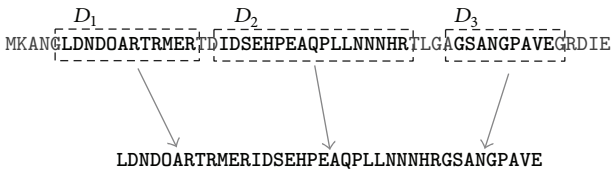

FIGURE 2: Illustration of restricting an amino acid sequence to which the spectrum kernel is applied to the domain regions.

where T indicates the total number of domains over all proteins and $M(D_m, P_i)$ indicates the number of domains identified as D_m in protein P_i.

2.8. Feature by Restriction of Spectrum Kernel to Domain Region (SPD). DN is based only on the number of constituent domains of each protein, while amino acid sequences of domains are also considered useful for inferring strength of PPI. Therefore, we propose a feature space mapping by restricting the application of the spectrum kernel [24] to domain regions (called SPD). Let $\mathscr{A}$ be the set of 21 alphabets representing 20 types of amino acids and others. Although we used the set of 20 alphabets to express 20 types of amino acids in the preliminary conference version [22], we add one alphabet to take the ambiguous amino acids such as X into consideration. Then, $\mathscr{A}^k$ $(k \geq 1)$ means the set of all strings with length k generated from $\mathscr{A}$. The k-spectrum kernel for sequences x and y is defined by

$$K_k(x, y) = \langle \Phi_k(x), \Phi_k(y) \rangle, \tag{7}$$

where $\Phi_k(x) = (\phi_s(x))_{s \in \mathscr{A}^k}$ and $\phi_s(x)$ indicates the number of times that s occurs in x. To make use of domain information, we restrict an amino acid sequence to which the k-spectrum kernel is applied to the domain regions. Figure 2 illustrates the restriction. In this example, the protein consists of domains D_1, D_2, D_3, and each domain region is surrounded by a square. Then, the subsequence in each domain is extracted, and all the subsequences in the protein are concatenated in the same order as domains. We apply the k-spectrum kernel to the concatenated sequence. Let $\phi_s^{(r)}(x)$ be the number of times that string s occurs in the sequence restricted to the domain regions in protein x in the above manner. The feature vector of SPD for proteins P_i and P_j is defined by

$$\begin{aligned} f_{ij}^{l} &= \phi_{s_l}^{(r)}(P_i) \quad \text{for } s_l \in \mathscr{A}^k, \\ f_{ij}^{(21^k+l)} &= \phi_{s_l}^{(r)}(P_j) \quad \text{for } s_l \in \mathscr{A}^k. \end{aligned} \tag{8}$$

It should be noted that $\phi_s^{(r)}$ for proteins having the same composition of domains can vary depending on the amino acid sequences of their proteins. That is, even if P_i and P_j have the same compositions as P_k and P_l, respectively, and the feature vector of DN for P_i and P_j is the same as that for P_k and P_l, then the feature vector of SPD for P_i and P_j can be different from that for P_k and P_l.

2.9. Support Vector Regression (SVR). To predict strengths of PPIs, we employ support vector regression (SVR) [25] with our proposed features. In the case of linear functions, SVR finds parameters w and b for $f(x) = \langle w, x \rangle + b$ by solving the following optimization problem:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|w\|^2 + C\sum_i \left(\xi_i + \xi_i'\right), \\ \text{subject to} \quad & y_i - \langle w, x_i \rangle - b \leq \epsilon + \xi_i, \\ & y_i - \langle w, x_i \rangle - b \geq -\epsilon - \xi_i', \\ & \xi_i \geq 0, \qquad \xi_i' \geq 0, \end{aligned} \tag{9}$$

where C and ϵ are positive constants and (x_i, y_i) is a training data. Here, the penalty is added only if the difference between $f(x_i)$ and y_i is larger than ϵ. In our problem, x_i means a protein pair, and y_i means the corresponding interaction strength.

2.10. Relevance Vector Machine (RVM). In this paper, we also employ relevance vector machine (RVM) [26] to predict strengths of PPIs. RVM is a sparse Bayesian model utilizing the same data-dependent kernel basis as the SVM. Its framework is almost the same as typical Bayesian linear regression. Given a training data $\{x_i, y_i\}_{i=0}^N$, the conditional probability of y given x is modeled as

$$p(y \mid x, w, \beta) = \mathcal{N}\left(y \mid w^T \phi(x), \beta^{-1}\right), \tag{10}$$

where $\beta = \sigma^2$ is noise parameter and $\phi(\cdot)$ is a typically nonlinear projection of input features. To obtain sparse solutions, in RVM framework, a prior weight distribution is modified so that a different variance parameter is assigned for each weight as

$$p(w \mid \alpha) = \prod_{i=0} \mathcal{N}\left(w_i \mid 0, \alpha_i^{-1}\right), \tag{11}$$

where $M = N + 1$ and $\alpha = (\alpha_1, \ldots, \alpha_M)^T$ is a hyperparameter. RVM finds hyperparameter α by maximizing the marginal likelihood $p(y \mid x, \alpha)$ via "evidence approximation." In the process of maximizing evidence, some α_i approach infinity and the corresponding w_i become zero. Thus, the basis function corresponding with these parameters can be removed, and it leads sparse models. In many cases, RVM performs better than SVM especially in regression problems.

3. Results and Discussion

3.1. Computational Experiments. To evaluate our proposed method, we conducted computational experiments and compared with the existing method, APM.

3.2. Data and Implementation. It is difficult to directly measure actual strengths of PPIs for many protein pairs by biological and physical experiments. Hence, we used WI-PHI dataset with 50000 protein pairs [23]. For each PPI, WI-PHI contains a weight that is considered to represent some reliability of the PPI and is calculated from several different

Table 1: Results of average RMSE for training and test data.

	$C = 1$		$C = 2$		$C = 5$	
	Training	Test	Training	Test	Training	Test
SVR + DN	0.10472	0.12573	0.10656	0.12600	0.09982	**0.12484**
RVM + DN	0.09210	0.12873	0.09178	0.12881	0.09474	0.12908
SVR + SPD ($k = 1$)	0.08819	0.12699	0.08080	0.12954	0.07927	0.12903
RVM + SPD ($k = 1$)	0.02848	0.12743	**0.01504**	0.12706	0.03276	0.12792
SVR + SPD ($k = 2$)	0.08891	0.12654	0.08188	0.12782	0.08117	0.12909
RVM + SPD ($k = 2$)	**0.02529**	**0.12470**	0.02301	**0.12476**	**0.02243**	0.12493
SVR + APM	0.06846	0.13112	0.06795	0.13247	0.06791	0.13277
RVM + APM	0.07052	0.13556	0.07037	0.13550	0.07032	0.13493
APM	Training = 0.06811, Test = 0.13517					

kinds of PPI datasets in some statistical manner to rank physical protein interactions. As strengths of PPIs, we used the value dividing the weight of PPI by the maximum weight for WI-PHI. We used dataset file "uniprot_sprot_fungi.dat.gz" downloaded from UniProt database [27] to get amino acid sequences, information of domain compositions, and domain regions in proteins. In this experiment, we used 1387 protein pairs that could be extracted from WI-PHI dataset with complete domain sequence via UniProt dataset. The extracted dataset contains 758 proteins and 327 domains. Since this dataset does not include protein pairs with interaction strength 0, we randomly selected 100 protein pairs that do not have any weights in the dataset and added them as protein pairs with strength 0. Thus, totally 1487 protein pairs were used in this experiment. We used "kernlab" package [28] for executing support vector regression and relevance vector machine and used the Laplacian kernel $K(x, y) = \exp(-\sigma\|x - y\|)$. The dataset and the source code implemented by R are available upon request.

To evaluate prediction accuracy, we calculated the root mean square error (RMSE) for each prediction. RMSE is a measure of differences between predicted values $\hat{y}_i$ and actually observed values y_i and is defined by

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}, \tag{12}$$

where N is the number of test data.

3.3. Results of Computational Experiments. We preformed threefold cross-validation, calculated the average RMSE, and compared with APM [17]. For APM method, strengths of PPIs are inferred based on APM scores for domain pairs that consist of target proteins. However it is not always possible to calculate APM scores for all domain pairs from training set. Therefore, as test set, we used only protein pairs that consist of domain pairs with APM scores calculated via training set. (In all cases, about 40% of protein pairs in test set were used.) For the Laplacian kernel employed in both SVR and RVM, we selected kernel parameter σ by fivefold cross-validation from candidate set $\sigma \in \{0.01, 0.02, \ldots, 0.1\}$. The parameter C for the regularization term in the Lagrange formulation is set to $C = 1, 2, 5$. Additionally, APM scores for each protein pair also can be used as input features. Therefore we also used APM scores as inputs for SVR and RVM and compared the model using APM scores with the model using our proposed features to confirm the usefulness of feature representation. Here, we used candidate set $\sigma \in \{3.0, 3.1, 3.2, \ldots, 9.0\}$ for kernel parameter σ of RVM + APM model because the model could not be trained with σ values smaller than 3. On the other hand, for σ of SVM + APM model, we used the same set as other models.

Table 1 shows the results of the average RMSE by SVR and RVM with our proposed features (DN and SPD of $k = 1, 2$) and APM score and by APM, for training and test datasets. For training set, the average RMSEs by RVM with SPD of $k = 2$ were smaller than those by APM and others. Moreover, for test set, all the average RMSEs by RVM with SPD and DN were smaller than those by APM. The results suggested that supervised regression methods, SVR and RVM, with domain based features are useful for prediction of PPI strengths. Taking all results together, the model by RVM with SPD of $k = 2$ was regarded as the best for prediction of PPI strengths.

Since the average RMSEs of SVR with APM for both training and test dataset were smaller than those of original APM, SVR has potential to improve prediction accuracies. By contrast, the average RMSEs of RVM with APM became larger than those of original APM, and all average RMSEs of the models with APM for test set were larger than those of the models with DN and SPD. Accordingly, the results suggested that prediction accuracies were enhanced by feature representation and SPD is especially useful among these feature representations for predicting strengths of PPIs. Although DN and SPD of $k = 1$ have 654 and 42 dimensions for each protein pair, respectively, the average RMSEs with SPD of $k = 1$ for training set were smaller than those with DN. It implies that information of amino acid sequence in domain regions is more informative comparing with information of domain compositions to make a model fit in with dataset.

In contrast, the RMSEs by SVR with DN were smaller than those by others in some cases of test set. Table 2 shows the numbers of relevance vectors and support vectors and the σ values selected by fivefold cross-validation in all cases. For the models with DN and APM scores, the numbers of relevance vectors were smaller than the numbers of support vectors. On the other hand, the numbers of relevance vectors were larger than the numbers of support vectors for the

Table 2: The number of relevance vectors (RVs) and support vectors (SVs) for each model with DN, SPD, and APM and the selected σ values for each fold.

		$C = 1$		$C = 2$		$C = 5$	
		SVR	RVM	SVR	RVM	SVR	RVM
		SVs (σ value)	RVs (σ value)	SVs (σ value)	RVs (σ value)	SVs (σ value)	RVs (σ value)
Fold 1	DN	271 (0.02)	113 (0.05)	271 (0.01)	123 (0.07)	308 (0.01)	74 (0.02)
	SPD ($k = 1$)	367 (0.01)	448 (0.02)	402 (0.02)	680 (0.05)	402 (0.02)	537 (0.03)
	SPD ($k = 2$)	392 (0.01)	502 (0.03)	409 (0.01)	628 (0.05)	421 (0.01)	628 (0.05)
	APM	362 (0.08)	4 (5.00)	361 (0.10)	6 (4.80)	357 (0.04)	6 (5.80)
Fold 2	DN	280 (0.02)	94 (0.08)	281 (0.01)	92 (0.09)	314 (0.01)	82 (0.04)
	SPD ($k = 1$)	408 (0.01)	617 (0.04)	453 (0.04)	706 (0.06)	411 (0.01)	545 (0.03)
	SPD ($k = 2$)	430 (0.01)	558 (0.04)	435 (0.01)	618 (0.05)	495 (0.04)	654 (0.06)
	APM	375 (0.10)	5 (6.50)	372 (0.10)	6 (6.90)	373 (0.04)	4 (5.50)
Fold 3	DN	321 (0.04)	107 (0.08)	289 (0.01)	107 (0.10)	330 (0.01)	107 (0.08)
	SPD ($k = 1$)	371 (0.01)	439 (0.02)	412 (0.03)	658 (0.05)	382 (0.01)	305 (0.01)
	SPD ($k = 2$)	387 (0.01)	625 (0.06)	418 (0.02)	529 (0.04)	398 (0.01)	529 (0.04)
	APM	368 (0.08)	3 (7.10)	368 (0.04)	3 (6.70)	372 (0.01)	5 (4.20)

models with SPD feature in spite of the fact that usually RVM provides a sparse model compared with SVR. In RVM framework, sparsity of model is caused by distributions of each weight; that is, the number of relevance vectors is influenced by values and variances of each dimension of features rather than by the number of dimensions of features. Actually, each dimension of SPD feature almost always has widely varying values. In contrast, DN feature has many zeros, and APM score is inferred from training dataset and thereby has similar distribution. Thus, it is considered that many weights corresponding to features in RVM model did not become zero and the RVM models with SPD feature tended to be complex and to overfit the training data.

4. Conclusions

For the prediction of strengths of PPIs, we proposed feature space mappings DN and SPD. DN is based on the number of domains in a protein. SPD is based on the spectrum kernel and defined using the amino acid subsequences in domain regions. In this work, we employed support vector regression (SVR) and relevance vector machine (RVM) with the Laplacian kernel and conducted threefold cross-validation using WI-PHI dataset. For both training and test dataset, the average RMSEs by RVM with SPD feature were smaller than those by APM. The results showed that machine learning methods with domain information outperformed existing association method that is based on the probabilistic model of PPIs and implied that the information of amino acid sequence is useful for prediction comparing with only information of domain compositions. However, the models with SPD feature tended to be complex and overfitted to the training data. Therefore, to further enhance the prediction accuracy, improving kernel functions combining physical characteristics of domains and amino acids might be helpful.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This work was partially supported by Grants-in-Aid nos. 22240009 and 24500361 from MEXT, Japan.

References

[1] J. R. Bock and D. A. Gough, "Predicting protein-protein interactions from primary structure," *Bioinformatics*, vol. 17, no. 5, pp. 455–460, 2001.

[2] Y. Guo, L. Yu, Z. Wen, and M. Li, "Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences," *Nucleic Acids Research*, vol. 36, no. 9, pp. 3025–3030, 2008.

[3] C. Yu, L. Chou, and D. T. Chang, "Predicting protein-protein interactions in unbalanced data using the primary structure of proteins," *BMC Bioinformatics*, vol. 11, article 167, 2010.

[4] J. Xia, X. Zhao, and D. Huang, "Predicting protein-protein interactions from protein sequences using meta predictor," *Amino Acids*, vol. 39, no. 5, pp. 1595–1599, 2010.

[5] J. Xia, K. Han, and D. Huang, "Sequence-based prediction of protein-protein interactions by means of rotation forest and autocorrelation descriptor," *Protein and Peptide Letters*, vol. 17, no. 1, pp. 137–145, 2010.

[6] S. J. Darnell, D. Page, and J. C. Mitchell, "An automated decision-tree approach to predicting protein interaction hot spots," *Proteins: Structure, Function, and Bioinformatics*, vol. 68, no. 4, pp. 813–823, 2007.

[7] K. Cho, D. Kim, and D. Lee, "A feature-based approach to modeling protein-protein interaction hot spots," *Nucleic Acids Research*, vol. 37, no. 8, pp. 2672–2687, 2009.

[8] J. Xia, X. Zhao, J. Song, and D. Huang, "APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility," *BMC Bioinformatics*, vol. 11, no. 1, article 174, 2010.

[9] O. Kuchaiev, M. Rašajski, D. J. Higham, and N. Pržulj, "Geometric de-noising of protein-protein interaction networks," *PLoS Computational Biology*, vol. 5, no. 8, Article ID e1000454, 2009.

[10] L. Zhu, Z. You, and D. Huang, "Increasing the reliability of protein-protein interaction networks via non-convex semantic embedding," *Neurocomputing*, vol. 121, pp. 99–107, 2013.

[11] S. Byrum, S. Smart, S. Larson, and A. Tackett, "Analysis of stable and transient protein-protein interactions," *Methods in Molecular Biology*, vol. 833, pp. 143–152, 2012.

[12] P. Uetz, L. Glot, G. Cagney et al., "A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*," *Nature*, vol. 403, no. 6770, pp. 623–627, 2000.

[13] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, "A comprehensive two-hybrid analysis to explore the yeast protein interactome," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 8, pp. 4569–4574, 2001.

[14] M. Hayashida, N. Ueda, and T. Akutsu, "Inferring strengths of protein-protein interactions from experimental data using linear programming," *Bioinformatics*, vol. 19, no. 2, pp. ii58–ii65, 2003.

[15] M. Hayashida, N. Ueda, and T. Akutsu, "A simple method for inferring strengths of protein-protein interactions," *Genome Informatics*, vol. 15, no. 1, pp. 56–68, 2004.

[16] E. Sprinzak and H. Margalit, "Correlated sequence-signatures as markers of protein-protein interaction," *Journal of Molecular Biology*, vol. 311, no. 4, pp. 681–692, 2001.

[17] L. Chen, L. Wu, Y. Wang, and X. Zhang, "Inferring-protein interactions from experimental data by association probabilistic method," *Proteins: Structure, Function and Genetics*, vol. 62, no. 4, pp. 833–837, 2006.

[18] R. Finn, J. Mistry, J. Tate et al., "The Pfam protein families database," *Nucleic Acids Research*, vol. 38, supplement 1, pp. D211–D222, 2009.

[19] S. Hunter, P. Jones, A. Mitchell et al., "InterPro in 2011: new developments in the family and domain prediction database," *Nucleic Acids Research*, vol. 40, no. 1, pp. D306–D312, 2012.

[20] J. Xia, B. Wang, and D. Huang, "Inferring strengths of protein-protein interaction using artificial neural network," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN '07)*, pp. 2471–2475, Orlando, Fla, USA, August 2007.

[21] M. Deng, S. Mehta, F. Sun, and T. Chen, "Inferring domain-domain interactions from protein-protein interactions," *Genome Research*, vol. 12, no. 10, pp. 1540–1548, 2002.

[22] Y. Sakuma, M. Kamada, M. Hayashida, and T. Akutsu, "Inferring strengths of protein-protein interactions using support vector regression," in *Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications*, 2013, http://world-comp.org/p2013/PDP2162.pdf.

[23] L. Kiemer, S. Costa, M. Ueffing, and G. Cesareni, "WI-PHI: a weighted yeast interactome enriched for direct physical interactions," *Proteomics*, vol. 7, no. 6, pp. 932–943, 2007.

[24] C. Leslie, E. Eskin, and W. Noble, "The spectrum Kernel: a string Kernel for SVM protein classi fi cation," in *Proceedings of Pacific Symposium on Biocomputing (PSB '02)*, pp. 564–575, Lihue, Hawaii, USA, January 2002.

[25] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, 1995.

[26] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, no. 3, pp. 211–244, 2001.

[27] T. U. Consortium, "Reorganizing the protein space at the universal protein resource (UniProt)," *Nucleic Acids Research*, vol. 40, no. D1, pp. D71–D75, 2012.

[28] A. Karatzoglou, K. Hornik, A. Smola, and A. Zeileis, "kernlab—an S4 package for kernel methods in R," *Journal of Statistical Software*, vol. 11, no. 9, pp. 1–20, 2004.